

Code Submissions

1. Github Repository: [Link](#)
2. requirements.txt

```
```
streamlit
scikit-learn
pandas
numpy
matplotlib
seaborn
xgboost
```
```

```

## README.md

### Breast Cancer Prediction & Analysis Pipeline

#### a. Problem Statement

The objective of this assignment is to implement an end-to-end Machine Learning classification pipeline. This involves training multiple classification models on a chosen dataset to predict a target variable, evaluating their performance using various metrics, and deploying the best-performing models via an interactive Streamlit web application. The goal is to demonstrate proficiency in the entire ML workflow: data selection, preprocessing, modeling, evaluation, and deployment.

#### b. Dataset Description

**Dataset Name:** Breast Cancer Wisconsin (Diagnostic) Dataset

**Source:** sklearn.datasets (originally from UCI Machine Learning Repository)

**Description:** Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

**Target:** Diagnosis (M = malignant, B = benign)

**Features:** 30 numeric features (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension).

**Instances:** 569 (Meets requirement of  $\geq 500$ )

**Feature Count:** 30 (Meets requirement of  $\geq 12$ )

#### c. Models Used & Comparison Table

The following 6 classification models were implemented and evaluated:

1. Logistic Regression
2. Decision Tree Classifier
3. K-Nearest Neighbor (kNN) Classifier
4. Naive Bayes Classifier (Gaussian)
5. Random Forest Classifier (Ensemble)
6. XGBoost Classifier (Ensemble)

### Evaluation Metrics Comparison

| ML Model Name       | Accuracy | AUC Score | Precision | Recall | F1 Score | MCC Score |
|---------------------|----------|-----------|-----------|--------|----------|-----------|
| Logistic Regression | 0.9825   | 0.9954    | 0.9861    | 0.9861 | 0.9861   | 0.9623    |
| Decision Tree       | 0.9123   | 0.9157    | 0.9559    | 0.9028 | 0.9286   | 0.8174    |
| kNN                 | 0.9561   | 0.9788    | 0.9589    | 0.9722 | 0.9655   | 0.9054    |
| Naive Bayes         | 0.9298   | 0.9868    | 0.9444    | 0.9444 | 0.9444   | 0.8492    |
| Random Forest       | 0.9561   | 0.9939    | 0.9589    | 0.9722 | 0.9655   | 0.9054    |
| XGBoost             | 0.9561   | 0.9901    | 0.9467    | 0.9861 | 0.9660   | 0.9058    |

### Observations about Model Performance

1. **Logistic Regression:** Performed exceptionally well, achieving the highest Accuracy (98.25%) and F1 Score (98.61%). This suggests the dataset is linearly separable to a high degree.
2. **Ensemble Models (Random Forest & XGBoost):** Both performed robustly with identical Accuracy (95.61%). They handle non-linear relationships well but were slightly outperformed by the simpler Logistic Regression on this test set.
3. **Decision Tree:** Had the lowest accuracy (91.23%) among the models, likely due to overfitting on the training data compared to the ensemble methods which mitigate this.
4. **Naive Bayes:** Performed reasonably well (93%) given its strong independence assumptions, showing that the features are likely independent enough for this model to be effective.
5. **kNN:** Achieved competitive results (95.61%), similar to the ensemble models, indicating that local neighborhood structures are preserving class information well.
6. **Overall:** All models achieved >90% accuracy, making them suitable for this task. Logistic Regression is the recommended model for this specific dataset and split due to its simplicity and superior performance.

# Streamlitt App

1. Application: [Link](#)
2. Application Screenshots:

**Model Selection**  
Choose a Classification Model  
Logistic Regression

**Input Features**  
Adjust the values below:

| Feature          | Value  |
|------------------|--------|
| mean radius      | 14.13  |
| mean texture     | 19.29  |
| mean perimeter   | 91.97  |
| mean area        | 654.89 |
| mean smoothness  | 0.10   |
| mean compactness | 0.10   |
| mean concavity   |        |

**Breast Cancer Prediction App**  
This app predicts whether a breast mass is benign or malignant using various Machine Learning models.

**Predict**

**Model Performance Comparison**

| ML Model Name         | Accuracy | AUC    | Precision | Recall | F1 Score | MCC    |
|-----------------------|----------|--------|-----------|--------|----------|--------|
| 0 Logistic Regression | 0.9825   | 0.9954 | 0.9861    | 0.9861 | 0.9861   | 0.9623 |
| 1 Decision Tree       | 0.9123   | 0.9157 | 0.9559    | 0.9028 | 0.9286   | 0.8174 |
| 2 KNN                 | 0.9561   | 0.9788 | 0.9589    | 0.9722 | 0.9655   | 0.9054 |
| 3 Naive Bayes         | 0.9298   | 0.9868 | 0.9444    | 0.9444 | 0.9444   | 0.8492 |
| 4 Random Forest       | 0.9561   | 0.9939 | 0.9589    | 0.9722 | 0.9655   | 0.9054 |
| 5 XGBoost             | 0.9561   | 0.9901 | 0.9467    | 0.9861 | 0.966    | 0.9058 |

**Current Model (Logistic Regression) Metrics:**

| ML Model Name         | Accuracy | AUC    | Precision | Recall | F1 Score | MCC    |
|-----------------------|----------|--------|-----------|--------|----------|--------|
| 0 Logistic Regression | 0.9825   | 0.9954 | 0.9861    | 0.9861 | 0.9861   | 0.9623 |

**Manage app**

**Model Selection**  
Choose a Classification Model  
Logistic Regression

**Input Features**  
Adjust the values below:

| Feature          | Value  |
|------------------|--------|
| mean radius      | 23.67  |
| mean texture     | 19.29  |
| mean perimeter   | 91.97  |
| mean area        | 654.89 |
| mean smoothness  | 0.10   |
| mean compactness | 0.10   |
| mean concavity   |        |

**Breast Cancer Prediction App**  
This app predicts whether a breast mass is benign or malignant using various Machine Learning models.

**Predict**

**Prediction Result**

**Malignant**

Confidence (Benign): 0.22  
Confidence (Malignant): 0.78

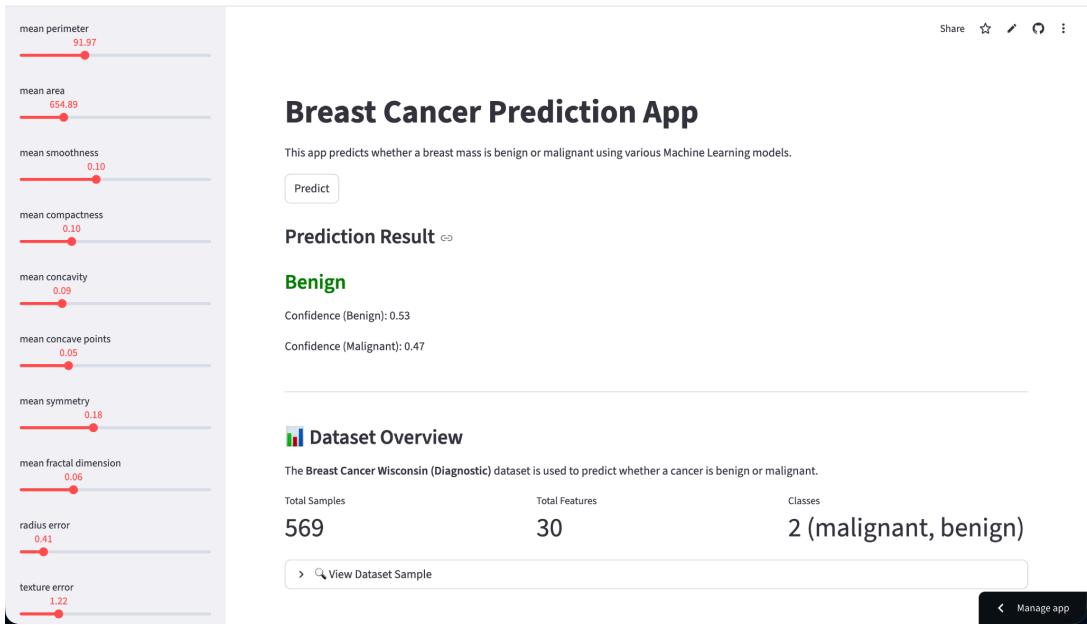
**Dataset Overview**

The Breast Cancer Wisconsin (Diagnostic) dataset is used to predict whether a cancer is benign or malignant.

| Total Samples | Total Features | Classes               |
|---------------|----------------|-----------------------|
| 569           | 30             | 2 (malignant, benign) |

**View Dataset Sample**

**Manage app**



## Lab Screenshots

### 1. Terminal Screenshots after training model

```

← → ⌂ argo-rdp.codeargo.net/web-rdp/#client/aS0wZTBmM2ZmNDlxZWY4MJU3YgBjAGpzB24?token=7BDD3A60C1B34763576F4EF5BD77FF526E7037AD... 🗃 ☆ 🌐 School ⋮
Applications Streamlit — Mozilla Fire... Terminal - cloud@2024d...
File Edit View Terminal Tabs Help Terminal - cloud@2024dc04109:~/projects/prediction-and-analysis-pipeline
(base) [cloud@2024dc04109 prediction-and-analysis-pipeline]$ python3 train_model.py
Loading Breast Cancer Wisconsin dataset...
Dataset Shape: (569, 30)
Features: 30
Instances: 569

Training and evaluating models...
Training Logistic Regression...
Training Decision Tree...
Training KNN...
Training Naive Bayes...
Training Random Forest...
Training XGBoost...
/home/cloud/anaconda3/lib/python3.12/site-packages/xgboost/training.py:200: UserWarning: [17:59:53] WARNING: /__w/xgboost/xgboost/src/learner.cc:782:
Parameters: { "use_label_encoder" } are not used.

bst.update(dtrain, iteration=i, fobj=obj)

Model Evaluation Metrics:
ML Model Name Accuracy AUC Precision Recall F1 Score MCC
Logistic Regression 0.982456 0.995370 0.986111 0.986111 0.986111 0.962302
Decision Tree 0.912281 0.915675 0.955882 0.902778 0.928571 0.817412
KNN 0.956140 0.978836 0.958904 0.972222 0.965517 0.905447
Naive Bayes 0.929825 0.986772 0.944444 0.944444 0.944444 0.849206
Random Forest 0.956140 0.993882 0.958904 0.972222 0.965517 0.905447
XGBoost 0.956140 0.990079 0.946667 0.986111 0.965986 0.905824

Models and scaler saved in 'model/' directory. Metrics saved to 'model_metrics.csv'.
(base) [cloud@2024dc04109 prediction-and-analysis-pipeline]$ █

```

## 2. Application Running on Localhost

The screenshot shows two instances of the Breast Cancer Prediction App running on a local machine. Both instances have identical URLs: `argo-rdp.codeargo.net/web-rdp/#/client/aS0wZTBmM2ZmNDIxZWY4MjU3YgBjAGpz24?token=7BDD3A60C1B34763576F4EF5BD77FF526E7037AD...`.

**Model Selection:** Choose a Classification Model: Logistic Regression.

**Input Features:** Adjust the values below:

- mean radius: 14.13
- mean texture: 19.29
- mean perimeter: 91.97
- mean area: 654.89
- mean smoothness: 0.10

**Prediction Result:**

**Benign**

Confidence (Benign): 0.53  
Confidence (Malignant): 0.47

---

**Dataset Overview:**

The Breast Cancer Wisconsin (Diagnostic) dataset is used to predict whether a cancer is benign or malignant.

| Total Samples | Total Features | Classes               |
|---------------|----------------|-----------------------|
| 569           | 30             | 2 (malignant, benign) |

The second instance of the application shows the same interface but with different input feature values and a different prediction result.

**Model Selection:** Choose a Classification Model: Logistic Regression.

**Input Features:** Adjust the values below:

- mean radius: 25.36
- mean texture: 19.29
- mean perimeter: 91.97
- mean area: 654.89
- mean smoothness: 0.10

**Prediction Result:**

**Malignant**

Confidence (Benign): 0.18  
Confidence (Malignant): 0.82

---

**Dataset Overview:**

The Breast Cancer Wisconsin (Diagnostic) dataset is used to predict whether a cancer is benign or malignant.

| Total Samples | Total Features | Classes               |
|---------------|----------------|-----------------------|
| 569           | 30             | 2 (malignant, benign) |